Detection of short size mutations and copy number alterations in ultra-deep targeted sequencing data

Valentina Boeva, Emmanuel Barillot,

Institut Curie, `valentina.boeva@curie.fr`

Jean-Francois Laes

OncoDNA, `jf.laes@oncodna.com`

The emergence of the amplicon sequencing technique, which followed whole exome sequencing, promises a revolution in cancer diagnostics and treatment. Amplicon sequencing consists of the PCR amplification of a limited number of the genomic regions of interest (amplicons) followed by high throughput sequencing [1]. Each amplicon often coincides with an exon; exons longer than the typical length of PCR reaction products may be covered by two or more amplicons (Figure 1).

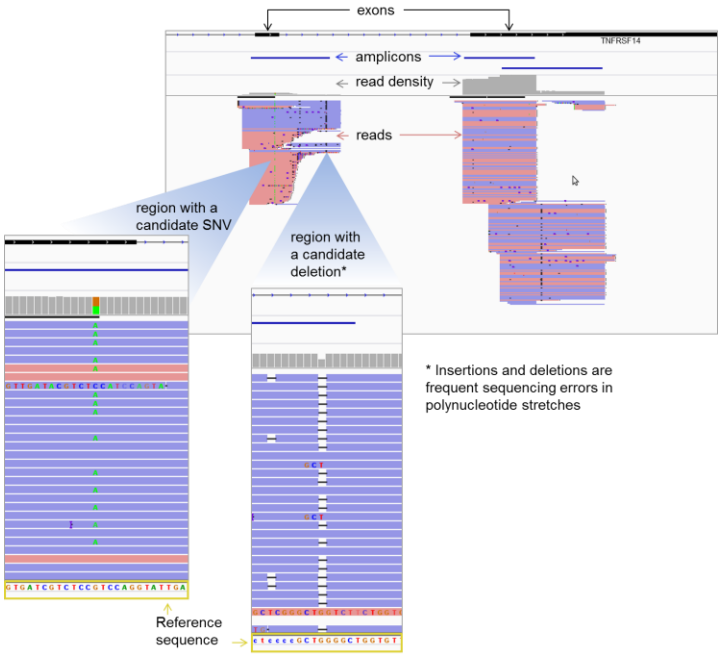

Figure 1. Visualization of the typical configuration of read mappings (with Integrative Genomics Viewer). Exons 7 and 8 of the *TNFRSF14* gene.

While relatively expensive exome sequencing consists of in-depth sequencing of nearly all the coding exons, the amplicon sequencing technique aims at sequencing a limited number of genes (from several dozen to several thousand exons) at an extremely low cost. The genes

included in a panel of amplicon sequencing (actionable genes) are genes that are often altered in different cancer types, and for whose alterations targeted therapies have been established or are in clinical development. For instance, the TargetRich™ CRX kit from Kailos Genetics assays such cancer-related genes as *BRAF*, *EGFR*, *FLT3*, *JAK2*, *KIT*, *KRAS*, *PIK3CA*, *PTEN*, *TP53* and *VEGFA*; the AmpliSeq™ Cancer Panel from Life Technologies targets 190 regions of interest in 46 well-characterized oncogenes and tumor suppressors.

Some actionable genes often undergo point mutation or exon deletions (e.g. *ALK*, *BRAF*) while others undergo amplification in copy number (e.g. *MYCN*, *ERBB2*) [2, 3]. Due to the exceedingly high read coverage of amplicon sequencing data, there is no methodological issue in the identification of clonal point mutations and small insertions or deletions (indels). However, how to reliably detect copy number changes, in particular gene deletions and amplifications, and identify subclonal mutations present in a very small proportion of tumor cells from amplicon sequencing data is still open to discussion.

Identification of copy number alterations (CNAs) of the actionable genes targeted by amplicon sequencing

Although there are several algorithms to detect CNAs in exome sequencing data [4-6], these approaches cannot be efficient when applied to amplicon sequencing data. First, amplicon sequencing targets fewer regions and thus provides less information than exome sequencing datasets (<10,000 exons vs >200,000 exons); consequently, data normalization can be less effective on amplicon sequencing data. Second, due to the different protocols used for library preparation, amplicon sequencing data can have various biases. Importantly, while for exome sequencing experiments an effort has been made to uniform exon coverage, amplicon sequencing technology emphasizes extremely high depth of coverage with less regard to coverage homogeneity.

Here we provide a solution to the challenging question of extracting CNAs from amplicon sequencing data by (i) defining a method to normalize read coverage with a small set of normal control samples and (ii) assigning statistical significance to putative CNAs resulting from the segmentation of normalized profiles. We validated the proposed method on (*i*) a high amplicon density dataset of 8 tumor samples for which array comparative genomic

hybridization (array CGH) profiles were available, (*ii*) a high amplicon density dataset of 30 ErbB2-positive ovarian cancer samples, and (*iii*) a low amplicon density dataset of 30 tumors, coupled with SNP array data. We show that the results obtained from the ONCOCNV method compare favorably with the results obtained from ADTEx [6] and NextGENe (http://www.softgenetics.com/NextGENe_013.html), respectively public-domain and commercial software designed to detect CNVs in whole exome sequencing data.

Detection of subclonal mutations in amplicon sequencing data

Given the extremely high coverage in amplicon sequencing datasets, the detection of germline or somatic mutations present in the majority of cells can be easily achieved using standard mutation calling tools [7]. However, we are often interested in detecting subclonal mutations, as the information about their presence can guide therapeutic choice. The task of detection of variants present in a very low proportion of cancer cells is obstructed by the presence of sequencing errors and possible read misalignments.

We propose a method, TargetZoom, to decrease the number of false positive predictions while keeping high the true positive rate. Our strategy includes several read post-processing steps and then several statistical tests to eliminate positions with possible sequencing or mapping errors.

The read post-processing consists of filtering out low mapping quality reads, low base quality positions within the reads (while taking into account polynucleotide stretches) and marking out mismatches corresponding to known SNPs. We group positions according to their *k*-mere context, where *k* is selected according to the total length of the targeted regions. Then, for a given position *i* in a context C_i, where we suspect a mutation, we apply filters 1-4 for check whether:

1. Read mapping quality for reads carrying the putative mutation is not worse than the read mapping quality of reads with the reference nucleotide at position *i*.
2. Base qualities of mutated bases are not worse than the base qualities of non-mutated bases for position *i*.
3. The proportion of reads with a mutation at position *i* is significantly higher than the

proportion of reads with the same mismatch at positions with similar context. By default, the method tests up to 10 different positions with a similar context.

4. The proportion of reads with a mutation at position *i* is significantly higher than the proportion of reads with this mutation in the control dataset. Several control datasets can be used.

At the end, we also discard positions that after all filters applied have more than two different mutation alleles.

The validation of TargetZoom was performed on the gold standard variant set for the reference individual NA12878 created by the NIST-led "Genome in a Bottle" Consortium (NIST-GIAB) [7].

References

1. C.Beadling et al. (2013) Combining Highly Multiplexed PCR with Semiconductor-Based Sequencing for Rapid Cancer Genotyping. *J. Mol. Diagn*., **15**:171–176.

2. L.A.Garraway and E.S.Lander (2013) Lessons from the Cancer Genome. *Cell*, **153**:17–37.

3. M.B.Small et al. (1987) Neoplastic transformation by the human gene N-myc. *Mol. Cell. Biol*., **7**:1638–1645.

4. V.Boeva et al. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**:268–269.

5. J.Li et al. (2012) CONTRA: Copy Number Analysis for Targeted Resequencing. *Bioinformatics*, **28**:1307–1313.

6. K.C.Amarasinghe et al. (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics*, **14**:S2.

7. H.Xu et al. (2014) Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*, **15**: 244.